DOCUMENT RESUME

ED 265 200 TM 860 056

AUTHOR Suchner, Robert W.
TITLE Evaluating College Courses: A Causal Analysis of Factors Linking Instructors' Activities and Students' Evaluations.
PUB DATE Apr 85
NOTE 57p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Evaluation Methods; Higher Education; *Questionnaires; *Student Evaluation of Teacher Performance; Teacher Effectiveness; Teacher Evaluation; Test Construction; *Test Reliability; *Test Validity; Undergraduate Students
IDENTIFIERS *Northern Illinois University

ABSTRACT

This paper is limited in scope to the preliminary questions that must be asked about a single instrument before one can use students' impressions of their teachers as part of any diagnostic or evaluation process. The described analysis is intended as illustrative of the kinds of questions that should be obtained to demonstrate the reliability and validity of students' evaluations and to justify using their evaluations of instruction as a basis for any decisions. The characteristics of an instrument developed by the Department of Sociology at Northern Illinois University to measure the relative teaching effectiveness of faculty in the department are examined. The questionnaire consists of 36 questions. Summary measures are included for students to give their overall evaluations of both the instructor and the course. Students are encouraged to write comments on the course. The construct validity of the instrument and factors that influence the students' summary evaluations are determined. The use of the questionnaire as a diagnostic instrument in providing information concerning how individual instructors might improve their teaching effectiveness and in making comparisons among the faculty for administrative decisions regarding salary, increments, tenure, promotions and retention is discussed. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED265200

EVALUATING COLLEGE COURSES:

A CAUSAL ANALYSIS OF FACTORS LINKING INSTRUCTORS' ACTIVITIES AND STUDENTS' EVALUATIONS

Robert W. Suchner

Department of Sociology
Northern Illinois University

April 1985

NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

U.S. DEPARTMENT OF EDUCATION

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Suchner, R. W.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented to the Division J session on the Interpretation and Use of Student Ratings of Instruction at the American Educational Research Association Annual Meetings, Chicago, Illinois, April 1, 1985.

TM 860 050

Getting faculty to thoughtfully consider and constructively use students' evaluations of their teaching is something like getting a child to taste something he/she just "knows" in advance won't taste good and "besides, it's not good for me." If you are not going to use a lot of authority you do not want to assert (and generate a great deal of resistance you do not want to deal with), you must be prepared to deal with an entire inventory of untested hypotheses (stated as "reasons") about why it would be better that you just go away and leave the matter alone.

This paper is the beginning of a test of some of the reasons faculty give for "just knowing" that students' evaluations of teaching are useless, a waste of time, and just something administrators do because they have nothing better to do. The paper is limited in scope to the preliminary questions that must be asked about a single instrument before one can entertain the use of students' impressions as part of any diagnostic or evaluation process. The analysis I will describe is intended as illustrative of the kinds of questions that should be asked and the kinds of information that should be obtained to demonstrate the reliability and validity of students' evaluations and to justify using their evaluations of instruction as a basis for any decisions.

THE RESEARCH QUESTIONS

There are essentially three questions that need to be answered before we can say that some behavior (such as giving answers to questions on a course evaluation form) is indicative of anything. The first question regards the construct validity of the behavior. In the case of students' evaluations of courses and instructors, this question can be phrased in three parts:

> What do students pay attention to when asked to evaluate courses and instructors?
>
> Do the answers they give reflect convergent validity in the sense that multiple statements tend to consistently reflect specific, interpretable concerns?
>
> Can and do students discriminate one concern from another, or do their answers to different questions simply reflect a "halo effect" of a single underlying impression?

If students pay attention to the questions they are asked, if their answers converge on a few underlying concerns that are supposed to be reflected in these questions, and if they demonstrate that they can and do discriminate among these concerns in the answers they give, we have the basis for saying that there is construct validity in their answers to our questions. The basic question then becomes: Have we

asked the kinds of questions that are useful in gathering information about the issues of concern to us as teachers and educational administrators?

The second question that should be answered is, in general: Which pieces of information provided by the students should be used for what purposes? Often we find ourselves collecting a lot of information and using only a summary measure. If we choose to do so with students' evaluations of instruction, we should ask the question:

> Which of the factors that students discriminate do they use to construct their overall summary evaluations?

If we find that the students' summary answers strongly reflect the concerns we have built into the questionnaire, and we find that these concerns are weighted in a reasonable way by the students, we may choose to use only their summary measures for certain purposes. Of course, if we find that students "over weight" some concerns and "under weight" others, we may want to either re-weight their responses to reflect our purposes and understandings, or we may want to augment the students' impressions with peer evaluations, self-reports, or other kinds of information (cf. Seldin, 1980). Clearly, if we find that the students' overall summaries are not at all reflective of the concerns we had hoped to measure, we would have a legitimate basis (and a reasonable set of arguments) to ignore the students' summary ratings.

Even if the students' discriminating responses represent a valid summary of teaching effectiveness, we may not want to use these summaries if they are contaminated by factors outside the control of the individual instructor. In particular, if we find that structural factors due to random influences or factors manipulable only by administrative decisions have appreciable effects on students' evaluations, we would either have to discount the differences in students' evaluations among instructors, or we would have to weight these evaluations by the differential effects of both positive and negative factors to which different instructors are exposed. The third research question arises from this concern and can be stated:

> What factors affect students' perceptions and evaluations of instructors and to what degree?

If we can rule out the irrelevant or non-manipulable "contaminants" in students' differential evaluations, we will be in a better position to "sell" our evaluation instrument and procedures to our reluctant faculty, and (quite consistent with the bases of their reluctance) we will be in a much

better position to make fair and reasonable sense of the students' evaluations in counseling both good and poor teachers on how to improve their teaching and in making administrative decisions both about how to better use the teaching talents different individuals display and who to retain, reward, and promote at least in part on the basis of evidence of differential teaching abilities and performance.

SOME RELEVANT LITERATURE

Unbeknownst to many faculty who object to the use of students' evaluations, there is a considerable research literature that demonstrates the general validity of students' answers to questions properly constructed and carefully administered and a growing literature that demonstrates that students can and do use appropriate bases for making judgments about the relative effectiveness of different instructors (see, for instance, reviews by Feldman, 1977, 1979). There is also a considerable literature that demonstrates the kinds of contaminants that need to be taken into account in making sense of students' differential evaluations (see, for instance, Miller, 1972, 1974; Seldin, 1980). In any instance of the application of a particular instrument and set of procedures, however, it is important to remember that what is true in general may not be realized in particular.

I will not provide an exhaustive review of these literatures in this particular paper. I will, however, use the reviews cited above and some more recent research by Marsh and his colleagues (Marsh, Overall & Kesler, 1979; Marsh, 1980; Marsh & Overall, 1981; Marsh, 1982; Marsh & Hocevar, 1984) and by Hoffman (1978) to suggest the bases on which we should expect our particular instrument to work and the problems we might confront in using it for diagnostic as well as administrative purposes.

CONSTRUCT VALIDITY OF STUDENTS' EVALUATIONS OF INSTRUCTION

To assess the construct validity of an instrument, one must state what it is one expects or hopes to measure. Seldin (1980: 159) remarks:

> Each source of information (student, colleague, administrator, self-assessment) offers important but limited insights. .... From students (one can reasonably expect to obtain): an appraisal of classroom teaching skills and the structure of the course (materials, exams, papers), work load, course difficulty level, professor-class interaction, professor-individual student

> interaction, student advising, organization, displayed interest in teaching, clarity of presentation, dynamism, enthusiasm.

Miller (1975: 31-32) argues in a similar vein:

> Students are in the best position to judge whether course objectives are clear and the course is well organized, whether the instructor explains clearly, allows for dissent, is patient, is interested in students and how he compares with other instructors with whom they have taken courses.

The instrument with which we will be working in this paper taps many of these dimensions.(1,2) Just as importantly, the instrument does not include items intended to collect information that the literature indicates students should not be expected to have or be able to report in a valid manner. The instrument is subsectioned into items dealing the "teaching skill," "my reaction," "course structure," "examination quality," and some other comments regarding the amount of work required, the amount of difficulty the student had with the work in the course, and two summary ratings -- one of the instructor and one of the course. (See Appendix.)

Each of these constructs has been shown by others to be measurable through students' evaluations. Seldin (1980) reports work by Guthrie (1954), Lovell & Haner (1955), Remmers (1959), Costin (1968) and Murray (1973) that demonstrate stability in students' evaluations on a number of these dimensions in the range of .70 to .90 test-retest reliabilities. Spencer (1968), Spencer and Aleamoni (1970), Harvey and Barker (1970), and Costin, Greenaugh, and Menges (1971) report internal consistency reliabilities ranging from .88 to .93. Feldman (1976: 243) has observed, following an extensive review of the literature: "(C)ollege students are only moderately consistent in rating their teachers and courses, although these modest interrater associations do produce substantial reliabilities for composite ratings when the ratings of at least 20 to 25 students in a class are averaged together." Similarly, the ERIC Clearinghouse of Higher Education Report 10 (1971: 10) concluded, after reviewing the research to that date: "Typical measures of split-half reliability consistently show that the instruments employed in student ratings have a high reliability. They have been checked and rechecked in different studies." Feldman (1977), Marsh (1982), and Hoffman (1978) report similar findings for reliability of the measures of the various constructs included in different instruments, and Miller (1974: 31) concluded: "The

evidence is clear and consistent....We can count on the reliability of student rating scale(s)...." if they are carefully constructed and administered as the numerous instruments he reviewed.

Regarding the validity of both component constructs and summary measures, there seems to be consistent evidence that students can and do discriminate among the constructs they are asked to report, and that multiple questions designed to measure the same construct can be developed that evidence convergence on these constructs and agreement with other forms of evaluation. Comparing students and peers' evaluations, Seldin (1980), for instance, reports work by Maslow and Zimmerman (1956), Costin (1966), Gaff (1973), Touq and Feldhusen (1973) that indicate appropriately positive correlations, given that faculty peers are often rating colleagues on a somewhat different basis than the students are asked to do. A particularly informative study on the correspondence between students' reports and others' direct observations, cited by Seldin (1980), was done by Touq, Feldhusen, and Halstead (1973). Two trained observers' recorded interaction data in 18 classes taught by 18 different instructors. Their observations were in considerable agreement with the reports of the 488 students' evaluating their instructors in these same classes.

Feldman's (1976) extensive review of studies entitled "The Superior College Teacher from the Students' View" demonstrated quite convincingly that numerous component constructs can and are consistently measurable using students' reports. Among the constructs included in this list of "measurables" were: stimulation of interest, clarity and understandability, knowledge of the subject matter, preparation for and organization of the course, and enthusiasm for the subject matter and for teaching. The research reviewed by Feldman indicates that these constructs are used by students to inform their overall evaluations of the "superior college teacher," while other constructs that might be some considered to be less relevant to the the evaluation of teaching, per se -- such as friendliness, concern and respect for students, helpfulness and availability to students, and openness to others' opinions -- while respected by students as characteristics of the superior teacher (and, one presumes, of people in general!) are not particularly predictive of students' overall evaluations of teaching effectiveness. In other words, given the opportunity to "bias" their evaluations toward individuals who appeared more gracious and easier to get along with, students tend to use what would probably be considered the more centrally relevant dimensions of teaching effectiveness to form their evaluations. Such a consistent finding is, indeed, strong evidence for the construct validity of students' evaluations of teaching effectiveness.

In a series of more recent studies, Marsh and his colleagues (Marsh, Overall & Kesler, 1979; Marsh & Overall, 1981; Marsh, 1982; Marsh & Hocevar, 1984) and Hoffman (1978) have provided extensive and convincing evidence that what students are discriminating are, in fact, the dimensions they are asked to report. In 1979, Marsh, Overall, and Kesler reported a study on 190 classes comparing students' ratings with instructors' self-ratings. They found that students' and self-ratings agreed to a considerable extent, not only on the overall evaluations but also on nine component factors. In that same year, Braskamp, Caulley, and Costin (1979) reported a study on 17 instructors and found that their self-ratings corresponded quite highly to their students' ratings of them on five components plus a global measure.

More recently, Marsh and Overall (1981) studied the correspondence between ratings of professors at the end of the semester and ratings made by students one year following graduation. They found a correlation of r=.83 and essential stability of means across this time period. They also found that these results held across subject matter and whether the course in question was an undergraduate or graduate course.

In 1981, Marsh reported a study of 341 pairs of courses in which the same instructor taught the same course on two separate occasions. The mean reliability of student ratings within each class was r=.93, while the mean correlation of ratings across the same classes taught by the same instructor was r=.71. It is interesting to note that some of the lack of correlation between semesters was predictable from the fact that faculty members tended to get somewhat higher ratings for courses they had taught at least once before.

In 1984, Marsh and Hocevar reported an extensive study involving 316 instructors and 31,322 students. Courses and instructors were selected such that each instructor had taught the same course at least four times over a period of four years. Marsh and Hocevar considered the construct validity and stability of class average ratings for nine dimensions: (1) learning/value of course, (2) enthusiasm of the instructor, (3) organization of the course, (4) group interaction (e.g., encouragement of discussion), (5) instructor's rapport with students, (6) breadth of coverage of related subjects, (8) examinations and grading, (8) assignments, and (9) workload/course difficulty. Using confirmatory factor analytic techniques, Marsh and Hocevar found convergent validity (across classes for the same instructor) of r=.68, very little "halo effect" (as indicated by correlated errors across traits within administrations), and considerable replication of the entire nine factor structure of the instrument across courses of different types taught by different instructors.

These findings should lay to rest any fears that, in general, students' evaluations of instructors do not replicate across semesters (i.e., are extremely susceptible to minor changes in presentation or chance factors outside the control of the instructor) or do not consistently measure the constructs they are intended to measure. It still remains to be demonstrated, however, whether any <u>particular</u> instrument (e.g., one newly constructed) has these admirable characteristics.

RESEARCH ON "CONTAMINANTS" AFFECTING STUDENTS' EVALUATIONS

Regarding our third research question -- identification of factors that "contaminate" students' evaluations and make it difficult to compare the teaching effectiveness of instructors teaching under different conditions -- the research consistently demonstrates that (a) such contaminants do exist, but (b) they tend to have little effect relative to the influence of the constructs of interest and (c) their effects are sometimes opposite to those generally expected.

Some of the expected contaminants studied in the literature are (1) class size, (2) level of the course in the curriculum, (3) whether the course is elective or required, (4) the time of day the course is given, (5) the composition of the students in the class (e.g., proportion freshmen, etc.), (6) the status of the student doing the rating (i.e., freshman, sophomore, etc.), (7) thegrade-point-average of the student, (8) the grade the student expects to receive in the class, (9) the student's prior interest in the subject matter, (10) the extent to which the course influenced the student's interest, and (11) a number of factors associated with how the evaluation questionnaire is administered (e.g., maintaining students' anonymity, with anouncements about how the information is to be used, with the professor present or absent).

Taking the last set of factors first, Feldman (1979) reports that a number of studies have shown very little difference in evaluations using distinctly different administrative procedures. Feldman (1979: 149) summarizes these studies by saying:

> (C)ollege students' ratings of their teachers and courses are somewhat higher when students remain anonymous rather than identifying themselves, when the purported use of the ratings is an official or administrative one for use in salary, promotion, or tenure considerations rather than otherwise, and when the instructor is present rather than absent during the rating session. (The

> differences between each of these contrasted circumstances are usually rather small and do not inevitably appear across studies.)...From limited evidence, the exact timing (or occasion) of evaluation appears not to be important in ratings. Variability in sampling procedures, as it affects the composition of students available to complete the rating forms, may or may not turn out to be a generally important element in ratings (as directly relevant data are collected).

Even though the literature reports little effect of these factors, it should be noted that the student evaluations to be analyzed in this paper were collected in the last quarter of each semester, the students remained anonymous to the professors in completing the questionnaire (both individually and collectively, in the sense that the students were told that their professor would not receive the results of their evaluations until after he/she had turned in the students' grades), and the professors were not present during the administration of the questionnaire (by university policy). Since there was no known variance on any of these factors, they will not be investigated in the following analysis.

Regarding a number of the "structural" factors that affect the context within which instruction was provided, the literature is quite clear that these factors do affect students' evaluations, but students appear to take such "obstacles" (e.g., size of class) into account and adjust their evaluations accordingly. Hence, the effect of structural differences across professors is somewhat minimal. For instance, Seldin (1980) reports that the findings on the effect of class size are mixed and never strong.

> Lovell and Haner (1975) found that instructors of classes larger than 30 students received lower ratings than instructors of smaller classes. McDaniel and Feldhusen (1970) reported the same finding. Other researchers, however, have found little or no correlation. Goodhartz (1948) found that instructors at Brooklyn College of classes of less than 20 students were not rated more favorably than instructors of larger classes. No significance was reported by Solomon (1966) between class size and ratings. An interesting curvilinear relationship was found by Gage (1961) who reported that teachers with 30-39 classroom students received lower ratings than teachers with more or fewer students. Centra (1976) reported a similar finding: instructors of classes under 15 students received exceptionally high ratings, but

> beyond that class size the relationship between size and ratings was minimal. In classes in 100 students, a slight elevation in student ratings was noted. (Seldin, 1980: 41-42)

Feldman's (1979) review finds the same lack of any strong relationship between class size and students' ratings, and interestingly Marsh, Overall, and Kesler (1979) report replicating the curvilinear relationship reported earlier by Gage (1966) and Centra (1976), though again, the relationship with students' ratings were very small except on the component of Group Interaction (i.e., discussion). Marsh, Overall, and Kesler (1979) also report that even this small relationship seemed to be less do to any systematic effect on students' ratings than it was a reflection of other "demographic" variables (e.g., percent of freshman and sophomore students enrolled).

Concerning other kinds of "structural factors," Feldman (1979) reports that correlations between students' ratings and the level of the course in the curriculum, whether the course is elective or required, and the subject matter of the course each are small but statistically significant. Level of course and electivity both have small bivariate relationships with students' ratings reported in numerous studies, while teachers in the humanities, the fine arts, and language appear to get somewhat higher ratings than instructors in other subject areas.

Regarding the students' own characteristics, which might vary across classes and "contaminate" comparisons, the literature indicates that the strongest factors are those associated with educationally relevant variables affecting the motivation and ability of a student to learn, while other factors that might introduce more capricious effects on the students' ratings appear to have little or no effect. For instance, Hoffman (1978) reports low bivariate correlations between students' ratings and expected grade, prior interest, and "outcome interest" (i.e., the increase in interest during the course). Hoffman reports no association, however, between the grade-point average of the student, his/her SAT math or verbal scores and the students' ratings of the instructors in mathematics and speech classes. where the general ability, math, and verbal facilities indicated by these measures would seem most relevant. Similarly, Hildebrand, Wilson, and Dienst (1971) and Kohlan (1973) report no appreciable effect of students' GPA's on their evaluations of the instructor, and Feldman concludes his review of the effects of prior interest in the subject matter by saying there is "no evidence one way or the other" that prior interest influences students' ratings (Feldman, 1976: 70).

Possibly the most comprehensive study that has been done of potential "contaminants" was reported by Marsh in 1980 and tends to confirm the results reported above. Marsh (1980) studied 511 undergraduate courses taught by 221 different instructors and analyzed the separate and combined effects of 16 background variables:

1. Prior interest in the subject
2. Workload/difficulty of the course
3. Overall GPA of the students
4. Class size
5. Academic rank of the instructor
6. Percent of the class majoring in the subject
7. Expected grade
8. Reason for taking the course (five measures)
9. Class level in the curriculum (four measures)

The results of Marsh's study indicate that (1) less than five percent of the variance of any component score on the students' evaluation was associated with any single background variable, (2) there was little curvilinearity in these bivariate relationships (which would mask association measured against a linear model), and (3) the proportion of variance explained by all the background factors combined was only 14 percent. The background variables evidencing the strongest associations with students' ratings of the instructor were (a) prior interest in the subject, (b) expected grade, (c) the work load, and (d) that the course was elective rather than required. All four of these variables were positively associated with the students' ratings, and a path analysis that specified prior interest as preceding grading expectations showed that prior interest accounted for approximately one-third of the association between expected grade and the overall rating.

These results, again, suggest that one is not going to get very far using either classroom context variables or differences in students' characteristics to account for differences among the evaluations instructors receive from their students. Stated differently, while there may be some structural conditions that should be taken into account in making comparisons between or among specific instructors' ratings, in general these kinds of variables represent neither "obstacles" to obtaining favorable student evaluations nor "contaminants" in making comparisons of the students' evaluations of different instructors or different classes.

APPLICATION TO A "NEW" INSTRUMENT

The literature considered above tends to show that, if students are asked questions relevant to what a given faculty consider important in evaluating teaching

effectiveness, the students' responses can often be shown to be both stable and predictive over time, internally consistent, discriminating, and generally uncontaminated by factors irrelevant to making comparisons among instructors or across classes. To be able to say these things about student evaluation instruments in general, however, is not to be able to attribute these characteristics to any particular instrument that might be developed by a particular individual or a particular faculty.

In the analysis that follows, I will examine the characteristics of an instrument developed by the Department of Sociology at Northern Illinois University to measure the relative teaching effectiveness of faculty in the department. In this preliminary analysis, I will attempt to establish the construct validity of the instrument and determine what factors influence the students' summary evaluations. I will then suggest how this instrument could be used as a diagnostic instrument in providing information about how individual instructors might improve their teaching effectiveness and how the instrument should be used in making comparisons among the faculty for administrative decisions regarding salary increments, tenure, promotion, and retention.

The Instrument

The instrument under consideration was developed about a decade ago following numerous discussions of the qualities members of the faculty considered important to consider in evaluating their own and others teaching effectiveness. The questionnaire consists of 36 questions, 24 of which are answered on a Likert-type scale. (See Appendix for the wording of questions used in the present analysis.) Summary measures are included for students to give their overall evaluations of both the instructor and the course (Q35 and Q36 in the Appendix). Students are encouraged to write any other comments they would like to make about the course on a separate sheet that is provided for this purpose.

Administrative Procedures

By university policy, the questionnaire is administered in the last fourth of each semester to all undergraduate courses with enrollments greater than five students. Students are informed that their responses to the questionnaire will be used as one piece of information in considerations of salary, tenure, and promotion of the faculty member they are rating and will be available to faculty for revising their courses. Students complete the questionnaire anonymously, having been

told that their instructor will not see the results until after he/she has handed in the grades for the semester. The instructor is, by university policy, expected to leave the room during the administration of the questionnaire.

The Constructs Intended to be Measured

The questionnaire is subsectioned into six parts:

1. General Information
2. Teaching Skill
3. My Reaction
4. Course Structure
5. Examination Quality
6. Other Comments and Summary Comments

Within each section are questions that would be expected to tap different aspects of the broader concept. For instance, under the section on Teaching Skill are two questions that are somewhat more general than the others and might be thought of as primarily reflecting the perceived preparedness of the instructor for each class period (Q9: Was well prepared for each class and Q6: Used class time well). Likewise, question Q15 (Presented too much material for the available time) might be thought of as relating as much to the amount of material covered in the course as to the instructor's skill in presenting the material.

CONFIRMATORY FACTOR ANALYSES OF THE STUDENTS' RESPONSES

Considering the various aspects of each more general concept, a confirmatory factor analysis was done to determine how well these items converged to a set of interpretable underlying constructs and how much the students discriminated among these constructs in their perceptions. The analysis was done using data from 3589 students providing complete data on these items in sociology classes between Fall 1981 and Spring 1984. These 3589 students represent a systematic random half-sample of each section taught by 27 faculty members who were assigned undergraduate classes in at least one consecutive Fall-Spring sequence during that time period. The analysis is restricted to this set of faculty to allow across-semester analyses to be done on the same data set after inspection of these preliminary results. The half-sample within each section was constructed to allow subsequent crossvalidation of the results of these analyses.

While the students, of course, provided all the information at the same time, a causal specification was anticipated in which answers to three distinct kinds of items might be thought of as representing causally linked constructs. Question Q13 asks the student to report whether

the instructor's "speech is adequate for teaching (volume, intelligibility, etc.)." This item will be taken in the following causal analysis as a subjective report of the student's assessment of the instructor's speaking ability which will be treated as a causally prior characteristic of the instructor and not included in the factor analytic models. Questions Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q15, Q19, Q20, Q21, Q22, Q23, Q24, and Q25 will be factor analyzed as a set of items all dealing with the students' perceptions of different aspects of the instructor's presentation in the classroom. Q14, Q16, Q17, Q18, and Q33 will be treated as the students' reports of their responses to the instructor's presentation and will, consequently, be factored separately and treated as causally a consequence of their perceptions of the classroom presentations.

Classroom Presentation

Dealing with the classroom presentation items first, the items were split into six sets dealing with what were expected to be six distinguishable characteristics. (See Table 1.) Confirmatory factor analysis (using LISREL VI, Joreskog & Sorbom, 1984; and SPSS-X, Nie, et al., 1983). was employed to estimate the convergence of the students' responses on these constructs. Perfect simple structure was specified such that each item loaded entirely on one and only one factor. The correlations among the factors were set free to be estimated, specifying the assumption that factors are generally correlated in the eyes of the students.

The standardized solution is given in Table 1. Inspection of alternative models indicated that each item loaded the highest on the a priori specified factor it was supposed to measure. The fit of the model can be assessed by the Goodness-of-Fit Index of .94 and the Adjusted Goodness-of-Fit Index of .91 (Joreskog & Sorbom, 1984).(3) Both indicate a reasonably good fit between the specified theoretical structure and the observed intercorrelations among the responses.

The six factors appear clearly interpretable, and each seems to represent considerable convergence of the students' responses to different items intended to measure essentially the same constructs. One can interpret Factor One either quite concretely as an indication of the students' judgments about how well the class time was used or in a somewhat more abstract manner focusing on the inferred reason the students thought the class time was used well (i.e., that the instructor was well prepared for each class). In either case, it seems that Factor One represents the students' general impressionsabout how well the class was run and how conscientious the instructor was in preparing for each class.

Factor Two taps a number of distinct aspects of the actual presentation of the material and is labelled to indicate an inference about the skill of the instructor in his/her presentations. It is not expected that these items would be perfectly correlated or would, equivalently, all perfectly measure a single unitary construct. Rather, they represent different kinds of things that different instructors will use to present his/her course material in a clear and organized fashion. For instance, it seems reasonable that question Q7 would load the strongest on this factor if the point of being skilled in the presentation of a course is to make clear the logical relationships among the different aspects of the course. Each of the other items on the factor appear to be ways individual instructors may go about making their presentations clear. In general, then, these items appear to measure the "skill" of the instructor in presenting the class in a clear and organized fashion.

Factor Three is a single item factor set to be equivalent to Q11. This single item seemed to be the only one which directly measured the attitude of the instructor toward the subject matter and the course. Other items are correlated with perceptions of the instructor's attitude (see the correlations among the factors, below), but each of these items appeared to more concretely describe the bases for inferences about the instructor's attitude than to provide direct measures of his/her attitude.

Factor Four is measured by two items (Q12 and Q21) that seem to reflect the instructor's patience with and openness to students in class. It is clear that these two items go together in the students' impressions, and it is probable that students are using the perceived patience of the instructor and the reactions of both themselves and others to assess a more general construct regarding the general atmosphere of the class.

Factor Five picks up variance from both the presentation of the material in class and the total amount of work that was required in the class. It is, of course, possible that instructors who require a lot of work from the students and whose syllabi cover a lot of material outside of class nevertheless do not try to cover too much material in the class. In this sense, the three items do not all belong together, Q15 having face validity for a measure of skill in presentation, Q31 appearing to tap the amount of material actually covered, and Q32 allowing the student to report how much work this actually involved on his/her part. However, the fact that these items all appear together on a single factor strongly indicates that this possibility of distinctions among these constructs is not, in fact, realized in the view of the students.

Factor Six appears to measure the perceived fairness of the instructor. Grading, in the experience of the author,

is a primary concern for many students. When an instructor is perceived to make the requirements and grading system clear and when tests are perceived to include a reasonable sampling of the material presented in the course (as against being selective or "tricky"), students tend to assess the instructor as "fair." These three aspects of the students' perceptions, then. tend to go together, even though only one of the items specifically mentions the "fairness" of the instructor.

Before examining the correlations among these factors, it might be useful to consider how the items might have loaded on a set of alternatively defined factors. For simplicity in conceptualization and interpretation, a perfect simple structure has been specified in which each item loads entirely on a single factor. Examining the reductions in Chi-square (i.e., the increased goodness-of-fit) that would occur if we relaxed the simple structure criterion, we can gain additional insight into the bases for the students' evaluations.

Four items showed clear evidence of complex structure (as indicated by a reduction in Chi-square of greater than 100 if each was released holding all others as specified). Q7 (Presented material in a logical and organized fashion), which loads highest on the Skilled factor, also shares considerable variance with items assessing the "preparedness" of the instructor and his/her "fairness" in presenting the class. Students apparently quite reasonably entertain the theory that disorganized and illogical presentations result from lack of time spent preparing for the class, and they appear to use such evidence as a basis for assessing the fairness of the instructor to the students. Whether the instructor actually is prepared for class or not, some of the students' overall rating of the instructor may be influenced by this multi-faceted set of inferences relating what students expect from professors in the form of preparation for each class, with what the student perceives about the logic and organization of the presentations and the provision of a fair opportunity to learn (and to obtain a good grade).

A second item that appears to tap more than one of our six dimensions is Q25 (The grading in the course was done fairly). Q25 loads highest on Factor Six (Fairness); however, it also would load quite high on either "preparedness" or "skilled." Again, the complexity of this item appears to indicate that students attribute the lack of fairness in grading to a lack of preparedness for the class. Or, stated positively, students seem to recognize that grading a course in a fair manner takes time and effort in preparation. Students also seem to be saying that they think that fairness in grading is a skill that goes along with other skills in presenting a class. Neither of these opinions appear to me to be unreasonable or unexpected.

Q10 (Used instructional materials where appropriate....) loads highest on the Skilled dimension, but it also appears to tap the students' assessment of the enthusiasm (Attitude) the instructor has for the subject. Students seem to be saying here that the instructor who routinely walks into class, puts down his/her notes, and spends the entire class period lecturing without handouts or any other aids to communication is not likely to be very enthusiastic about the subject or about communicating the subject to the students. This is a lot to read into a single item, but it is interesting that this item, which has been ridiculed as irrelevant in most discussions by our faculty when considering the students responses, actually shows some variance associated with a rather important aspect of teaching (i.e., the enthusiasm for the subject one communicates to the student).

The last complex item to be considered here is Q19 (The instructor clearly explained objectives of the course). Q19 loads highest on the Skilled factor. However, it also appears to measure the students' assessments of the "fairness" of the instructor in presenting the course. Again, this may reflect students' concern with grading. Or, giving students the benefit of the doubt, it may represent their concern for having an opportunity to learn. In any case, instructors who do not make clear what their objectives are for the course are perceived by the students as not providing a fair basis for their own participation in the course. This assessment seems to be consistent with the general conclusion that students expect instructors to be skilled in their presentation of the course, and when they are not, students are likely to evaluate instructors as somewhat "unfair."

Correlations Among the Factors

The correlations among these six factors suggest much of the same argument given above for the complexity of individual items. (See Table 2.) While students appear to discriminate somewhat between their perceptions of the "preparedness" and the "skill" of the instructor, these two dimensions are so highly intercorrelated that one might better consider them a single construct. We will not do so in this preliminary analysis, however, until after we have examined whether these two factors account for any independent variance in the students' overall evaluations (see below).

The other correlations in Table 2 are much lower but still reflect consider colinearity among at least five of the constructs. As we saw above with the individual items that might load on more than one factor, students tend to see the "skill" of the instructor as highly correlated with his/her preparedness for class, his/her attitude

(enthusiasm) for the subject, his/her patience with students in class, and the fairness of the class for the students. This cluster is quite consistent with the dimensions Feldman (1976) found defining the "Superior College Teacher" in previous research.

From the standpoint of alternative hypotheses faculty are sometimes heard to offer for (usually low) course evaluations, it is interesting to note that the perceived "workload" of the course is slightly negatively correlated with three dimensions of this "superior teacher" cluster and uncorrelated with the other two dimensions. These findings seem to suggest that there is at least some basis for the presumption that increasing the amount of work students are expected to do, and the amount of material presented in each class, is associated with lower assessments of the skill of the instructor and the judged fairness of the instructor in presenting the course, at least when the workload is perceived by the students as "excessive." It is also interesting to note, however, that the perception of the amount of work required is associated (negatively) with the "patience" of the instructor with students' questions in class and his/her perceived openness to questions and the expression of (possibly contrary) opinions. This correlation suggests that, in the eyes of students, the instructor who is particularly concerned with covering a lot of material is not necessarily the best teacher of the material he/she desires to cover. This interpretation is reinforced by the lack of correlation between the "work" factor and the students' perceptions of the instructor's preparedness or enthusiasm for the subject. Appearently students allow that the instructor who presents "too much" material in class is likely to be prepared and enthusiastic. They simply seem to be saying that presentation and requirement of an excessive amount of work gets in the way of effective interactions between students and professors and interferes with the fair presentation of the course.

Students' Responses to their Classes

Five items were interpreted as indicators of students self-reported responses to different aspects of the course (see Table 3). Three items measured the "usefulness" of the course to the student. Q16 (Found the course material useful for career preparation) is a rather straight-forward statement about the usefulness of the course if one takes the oft-stated view that students are in college today primarily to prepare for (or even just to become eligible for!) a job. The other items, however, reflect a different kind of "usefulness" more in line with traditional and often hoped for outcomes of a college education. Responses to these three items, then, represent a cluster of perceptions

that is broader than the most narrow interpretation that students find only the most "vocational" courses useful for their own purposes.

The other two items analyzed with this "usefulness" cluster are taken as representing single-item indicators of the students' "outcome interest" (c.f., Hoffman, 1978) and the perceived difficulty of the course for the student. It should be noted in passing that the item tapping the perceived difficulty of the course was written to distinguish what is perceived as a "heavy" course (as indicated by the "work" dimension) from the assessment of the level at which the material is presented relative to to background and ability of each student. The "Work" factor is intended to gather information about perceptions of the course, while the "Difficulty" factor is intended to gather information about the students' responses to the course.

The Goodness-of-Fit Index for the factor analysis of these five items is .96 and the Adjusted Goodness-of-Fit Index is .92.(4) Both indicate a quite close fit between the specified theoretical structure and the observed intercorrelations. However, again a better fit could be obtained by relaxing the restriction to simple structure. In this case, Q16 (Found the course material useful for career preparation) would load quite heavily on the "Interest" factor as well as the "Usefulness" factor. This makes quite good sense, but it also tends to confirm the notion that students put a great deal of weight on the the extent to which they perceive that a course is preparing them for a career. Apparently, the less connection students see between the relevance of a course for career preparation, the less they are likely to find their interest in the subject matter increased through participation in the course. It would be hopeful, but I am afraid somewhat premature at this point in our studies, to assume that the opposite is also true. That is, to the extent that instructors can use examples and materials of clear relevance to students' career interests, their perceived usefulness of the course and their perceived increase in interest in the subject matter is likely to increase.

The intercorrelations among these "response" items are given in Table 4. Perceived "Usefulness" and "Interest" are very highly correlated, as suggested by the analysis above. The perceived "Difficulty" of the course, like the "Work" factor in the previous factor analysis, is slightly negatively correlated with the other two factors. Again, this suggests that students who perceive the course as too difficult for their background and abilities are likely to find the course less interesting and less useful. While this is not a surprising finding, it does suggest that if students are to be kept interested in the subject matter and encouraged to see the relevance of what we are teaching, it

is important that we present the material in such a way that students do not find it too difficult to comprehend.

Conclusions Regarding Our First Research Question

Regarding our first research questions, we can say on the basis of the preceding analyses that (a) the items in the present questionnaire appear to tap the underlying dimensions they were intended to measure, (b) the students' responses to these items tend to converge to a great extent in the predicted manner, and (c) students' seem to be able to discriminate many of these dimensions from each other. The very high colinearity between "Preparedness" and "Skilled and that between "Usefulness" and "Interest" caution us that students do not find much to discriminate among these factors, and the complexity of a number of individual items suggests that students are using much of the same information to make multiple inferences about the characteristics of the course and their responses to it. In any consideration of these factors, or the individual items that make them up, we must then be careful to consider that when we ask students 36 different questions, we may not be getting 36 distinct pieces of information back.

WHAT FACTORS DO STUDENTS USE TO INFORM THEIR OVERALL EVALUATIONS?

Having determined something of the validity of students' perceptions of their courses and their responses to them, we can turn to the question of how students use this information to form an overall evaluation of the course and the instructor. As a first approximation to this question, we will examine the bivariate correlations between each factor and each overall evaluation (see Table 5).

The bivariate association is of interest because it helps us understand why and how people develop (usually implicit) theories about the determinants of high and low course and instructor evaluations. It is generally accepted that people develop notions about causality from observing correlations among events (cf. Heider, 1958). It is difficult to ask, however, that people in their everyday experiences "control for" covariates of the variables they are observing. In fact, it is my guess that most people, most of the time, observe and take into account only the bivariate associations among variables in informing their theories of behavior, while some of the time the more "complex thinkers" or more sesitive observers may take into account the first-order partials.

Having taken that brief tangent, we can examine Table 5 to see what most students and most faculty are likely to have "seen" in their multiple years of experience with the

association between the components of students' evaluations and their overall judgments. The data in Table 5 seem quite clear. All of the components of the students' evaluations, except their perceptions of the "workload" of the course and their "difficulty" with the course, are correlated quite highly with the students' summary ratings of the instructor. Consistent with Feldman's (1976) description of "The Superior College Teacher from the Students' View," students rate instructors higher as a positive function of their perceptions of his/her preparedness for class, skill in the classroom, enthusiasm for the subject, patience with students, fairness in presenting and grading the class, the perceived usefulness of the course, and the extent to which the class increased the student's individual interest in the subject. It is interesting to note that, even without controlling for these other components, students do not seem to rate instructors very differentially as a function of the workload the instructor requires or the difficulty the student has with the material.

Looking at the students' overall ratings of the course, we find a similar but not identical picture. In this case, students do seem to be somewhat "put off" when they perceive the workload as "excessive," and they similarly appear to rate the course somewhat less favorably when they have found it too difficult for their own background and abilities.

The fact that the components of the students' evaluations correlate somewhat differently with their overall ratings of the instructor and the course again attests to the fact that students are discriminating different aspects of the course and using their judgments of these distinct aspects as bases for making somewhat different statements about the instructor and the course. While it is not surprising that the overall ratings of instructor and course correlate quite highly, overlapping about 50 percent of their variance, it is interesting to note that this correlation is no higher than the correlations with a number of the component parts of the students' judgments, and that another 50 percent of the variance in the overall ratings is available to be explained by differential associations with the components and other factors.

To return briefly to our tangent into the individual psychology of perception and causal inference, the correlations in Table 5 suggest that students and faculty who have assumed that students' ratings of the course and the instructor are simply repetious reflections of the same judgment are about half right. However, those students and (in my experience more likely) faculty who "just know" that students overall ratings of the course and the instructor bear no relation to the students' perceptions of specific aspects of the course are almost certainly wrong.

ARE THERE "CONTAMINANTS" IN STUDENTS' OVERALL RATINGS?

To provide a proper test of the naive observer's (Heider, 1958) causal theory of students' evaluations, however, one must examine and be able to rule out the numerous alternative hypotheses (i.e., "reasons") faculty often are heard to present in explaining instructor and course ratings. To do this, we must explicitly state what these alternative hypotheses might be and identify the variables of purported importance in each.

In my experience, there are at least eight frequently heard and fervently believed alternative hypotheses.

1. THE STUDENT PREFERENCE HYPOTHESIS:

   Students like certain kinds of courses more than others. If you are assigned to teach the kinds of courses students don't like or find difficult, you will of course receive lower students' evaluations of your teaching.

2. THE EASIER TO TEACH HYPOTHESIS:

   Some courses are inherently easier to teach than others. It depends in part on the subject matter. If you are teaching subject "X," for instance, it is much harder to get good teaching evaluations than if you are teaching subject "Y." (Interestingly, it seems invariably that the instructor doing the speaking finds his/her own subject matter among the "inherently" most difficult to teach.)

3. THE CLASS STRUCTURE HYPOTHESIS:

   It's the structure of the class that matters. Obviously, you can keep students' attention and interest easier in some class situations than in others. That's why students "always" give better ratings in smaller classes, in classes in their major, in elective rather than required classes, in classes that meet in the middle of the day, and in shorter rather than longer classes.

4. THE TODAY'S STUDENTS DON'T SHARE OUR VALUES HYPOTHESIS:

Students aren't as serious as they used to be; they don't share or even understand the academic values of the university. If you try to do anything of value in the classroom, you get "marked down" by the present students. That's the reason that, even just over the last few years, the evaluations of the more serious professors have been going down.

5. THE "IN THE SPRING THEY'RE TIRED" HYPOTHESIS:

It is clearly evident that students are more tired iof school and "antsy" in the spring when the weather finally breaks in northern Illinois. By giving the course evaluations at the end of the semester, when students are the most likely to be looking forward to getting out for the summer, we exacerbate this effect. Consequently, if you are lucky enough not to have to teach in the spring, or even if you have smaller classes in the spring than in the fall, you are likely to get higher overall ratings from the students.

6. THE BETTER STUDEN[illegible]YPOTHESIS:

It's just the luck of the draw. If you get the better students -- the ones with the higher GPA's -- or you are teaching primarily juniors and seniors rather than freshmen and sophomores, you are very likely to get better ratings of your teaching.

7. THE ATTRACTION HYPOTHESIS:

It really comes down to whether the students like you or not. If they like you, they will like the class and will give you a good evaluation, whether or not they learn anything. If they don't like you, there is nothing you can do to get a good evaluation. We all know about "halo effects," and students' evaluations are just another example of a halo effect based on personal prejudices.

8. THE EASY GRADER HYPOTHESIS:

It's really none of the above. We all know that if you are an easy grader and students expect to get a higher grade from you than they normally get from others, you are a shoe-in for high teaching ratings.

A number of the variables called for by these hypotheses are measured either directly or indirectly with the present instrument. The following variables have been defined for subsequent use in the analysis of the causal structure implied in the above hypotheses (5):

LEVEL: The level of the course in the curriculum (freshman, sophomore, junior, or senior).

PCTMAJOR: The percentage of the class that are sociology majors.

NRATERS: The number of students providing usable ratings in each section of each course. (Used as a surrogate for class size, the latter having been demonstrated earlier (Suchner, 1984) to be highly correlated with the latter.)

THEORY, METHODS, ORGINST, SOCPSYCH, CRIMJUST: Five dummy variables indicating the subject matter of the class. (The sixth dummy variable: INTRO: becomes the comparison category and represents the introductory courses for each of the subject areas.)

STATUS: Whether the student is a freshman, sophomore, junior, or senior.

REQUIRED: Whether the class was required or elective for the student.

GPA and EXPGRADE: The student's reported grade point average from other classes to date and the student's reported expected grade for the present course. (Note: The evaluations were done in the last quarter, and often the last week, of each semester; hence, the students usually had a good basis for judging their expected grades in the present class.

MAJOR: Whether the student is a sociology major or not.

TERM: The semester (Fall or Spring) in which the class met.

YEAR: Year of the course (1981 through 1984).

Introducing these variables and one other -- Q13: Speech is adequate for teaching.... -- which is often posited as an alternative hypothesis without explicit indications whether the theorist considers this a controllable (i.e., manipulable) or uncontrollable "background" factor, we are ready to specify the causal model that might correspond, alternatively, to the hypotheses that suggest that is background factors (i.e., "contaminants") that account for students' evaluations versus the hypothesis put forth in this paper that students' perceptions of how the instructor presents the class account for the students' evaluations of the course and the instructor. (See Table 6 for the bivariate correlations of these background variables with the students' evaluations of the instructor and the course.)

The causal specification is quite simple but, as always, somewhat debatable. (See Figure 1.) The variables suggested by the alternative hypotheses are specified as exogenous to the causal system, being primarily under the control of administrative decisions affecting the assignment of courses to instructors and the scheduling of classes and otherwise due to chance factors affecting the assignment of students to classes. Variable Q13 (the judged adequacy of the instructor's speech) is specified as the first endogenous variable on the grounds that (a) it is a relatively constant factor that may influence students' perceptions of the presentation of any particular class but is unlikely to be influenced by these perceptions or actual changes in classroom presentation factors, and (b) it is a behavior directly observable and assessible by the student and is unlikely to be inferred on the basis of the students' other perceptions of the classroom presentation.(6) The six factors representing students' perceptions of the presentation of the course by the instructor are entered next, followed by the three factors representing the students' self-reported responses to the course. Finally, the students' overall, summary ratings of the course and the instructor are specified as causally the result of all the previously specified factors.

The block-recursive model that results corresponds quite closely to both the "effective student evaluator" hypothesis and the alternative "structural determinants" hypotheses and pits these two distinct theories against each other in a very explicit manner. If the "structural determinants" hypotheses are correct, estimating the parameters of this model should show that most of the variance in students' evaluations of both instructors and courses are

due to the direct effects of the exogenous variables. The relationships between the components of the students' evaluations, shown earlier to be quite large when these exogenous factors are not controlled, should be shown, for the most part, to be spurious relationships due to their common dependence on the specified structural factors. On the other hand, if the "Effective Student Evaluator" hypothesis is correct, we should find that the estimates of the parameters of this model show few, if any, direct effects of the exogenous variables on either the component evaluations or the summary student ratings, and that the components of the evaluations account for much of the variance in the summary ratings.

Estimates of the parameters of the model were obtained through ordinary least-squares linear regression using SPSS-X (Nie, et al., 1983). The assumptions of OLS regression were checked through examination of the residuals. Given the truncated range of the single-item indicators (based on five-point scales), the expected ceiling effects were found in the distributions of the residuals and the scatter diagrams of the partialed relationships. However, except for these ceiling effects, that would tend to lead to underestimates of associations, all the variables evidenced reasonable variances, and the distributions of the endogenous variables each approximate the normal curve. (See Table 7 for description of the central tendencies and dispersions).

Inspection of the partialed scatter plots revealed no indication of curvilinear relationships of any of the predictors with any of the endogenous variables. In particular, there was no evidence of any curvilinear relationship between size of class and any of the endogenous variables it might predict. Hence, there is no evidence in the present data to support Gage's (1961), Centra's (1976), and Marsh's (1980) findings that very small and very large classes are associated with higher ratings.(7)

The complete causal structure was estimated then trimmed to those factors which evidenced substantively important contributions.(8) The procedures used to trim the model were to subsection the blocks of the design into sets of comparable variables then inspect the step-down F-ratios for each subsection, removing all the variables in a subsection at once when all the variables, operating together, did not prove reliable at $p < .001$.

Next, models were examined eliminating one variable at a time when each proved nonsignificant beyond $p < .001$. In a final step, variables were removed, one at a time, if the path coefficients connecting them with other variables were all less than .10. For simplification of presentation, variables were also removed if they represented an "end point" in a causal chain that did not lead to an ultimate effect on the students' overall ratings of either their instructors or their courses. The resulting model, therefore, is restricted to variables that have demonstrable direct or indirect effects on the students' overall ratings. Variables with only spurious relationships have all been removed.

The resulting path diagram is given in Figure 2. Before discussing the causal effects evident in Figure 2., it would be instructive to examine which variables thought by the "structuralists" to be important causes of students' overall evaluations have been dropped out of the model by the above stated criteria.

Almost all of the exogenous variables have disappeared. The only remaining exogenous variables are Q13 (renamed "SPEECH" in the diagram), the subject field of sociological theory, and the students' expected grades in the present course. Removal of these exogenous variables totally refutes the "Class Structure" hypothesis, the "In the Spring They're Tired" hypothesis, and the "Today's Students Don't Share Our Values" hypothesis. None of the variables frequently mentioned as important structural factors to which faculty are assumed to be differentially exposed (e.g., size of class, level of class, required versus elective classes, proportion of students majoring in sociology, etc.) proved to contribute anything substantial, either directly or indirectly, to observed differences in the students' overall evaluations. And, neither the term in which the course was given nor the year showed any statistically significant direct or indirect effects on students evaluations. In fact, the data evidence remarkable stability between semesters and across years. Neither the weather nor systematic changes in the student population appear to account for any variance in students' evaluations of their courses or instructors.

Likewise, the "Student Preference" and "Easier to Teach" hypotheses have been all but eliminated. Except for the relatively small effect that indicates that students generally find "theory" courses to be less "useful" than courses in other subfields of the discipline, there appears to be little or no systematic differences in the ratings of instructors teaching in these somewhat distinct subfields. In particualr, it is interesting to note that "level of the course," "student status," and "introductory courses" (the comparison group for the subject field dummy variables) have

all dropped out of the model. Apparently it is not true (at least in the department under investigation) that instructors in introductory courses, those teaching mostly freshmen (or mostly seniors, depending on how the alternative hypothesis is stated), or those teaching lower (or upper) division courses are at any disadvantage when it comes time to consider the students' evaluations. In fact, the absence of all these exogenous variables provides rather strong evidence that much of the reasoning behind the "structural determinants" hypotheses is faulty and based on misperceptions of the causal process resulting in the students' overall evaluations.

There is some evidence, however, for some effects of what might be considered structural, or non-manipulable, variables. For instance, the student's expected grade in the course appears to affect his/her perceptions of the instructor's "patience" with students in the class, the "skill" of the instructor, and the "usefulness" of the course. These effects tend to be small relative to the effects of other variables in the model, but they apparently do exist, and they represent indirect paths through which differences in expected grades can "contaminate" students' overall evaluations. This finding is consistent with results reported by Holmes (1972) indicating that discrepancies between a student's GPA and his/her expected grade in a given course can effect the rating the student gives of the instructor in the course. Having controlled for the student's reported GPA in other courses to data, the reported "expected grade" represents just such a discrepancy, and the present findings tend to replicate Holmes'.

Does this mean that the "Better Student" and/or the "Easy Grader" hypotheses are supported? The fact that GPA has dropped out of the model appears to indicate that differences in the academic abilities of students do not affect differences in their overall ratings of instructors or courses. Hence, the general sense of the "Better Student" hypothesis does not appear to be supported. However, the present data are consistent with the more specific sense that the "better students" within the class (i.e., those who have "improved" the most over the prediction we would have made based on their overall GPA's) do, indirectly, rate both instructors and courses somewhat higher than the students who have not improved or who have done worse than expected. These data are also consistent, however, with the "Easy Grader" hypothesis.

How, then, do we choose between these two interpretations? Or, are we left with having no basis to choose between the somewhat optimistic statement that students who learn more (as indicated by higher than predicted grades) rate the course and the instructor higher and the more

cynical suggestion that, in order to get good evaluations, instructors simply need to give out a lot of high grades?

One basis on which to choose between these two points of view is to consider how expected grades impact on students' overall ratings. One path is through perceptions of the "patience" of the instructor. Is it possible that "patient" instructors are able to teach better, and that better teaching leads to higher grades for those students who will ask their questions? Another path is through the perceived "usefulness" of the course. Is it possible that students who come to see the course as "useful" for their purposes are more likely to learn more and "over-achieve" in a particular class? A third path is through "skillful." Is it possible that more skillful teachers, or at least those whose "skills" match the needs of the students (cf., Feldman, 1977), give the students the opportunity to learn more and exceed even their own expectations?

Each of these alternatives to the "alternative hypotheses" represent a re-specification of the model suggesting that the students' expected grades are an outcome rather than a determinant of the perception (and possibly the actuality) of patience, relevance, and skill in teaching. With cross-sectional data from students in multiple classes exposed to many different instructors, it seems impossible to definitively disentangle these interpretations. However, the general sense of the hypothesis is promising, and the consistency of this finding, both in the literature and in the pattern of effects in the present study, should encourage us to test the causal direction of this relationship experimentally or through repeated course evaluations within a single semester.

I have tried to deal in some detail with each of the alternative "structural determinants" hypotheses, except one, to allow each its chance to account for the observed data. The one I have excluded -- the "Attraction Hypothesis" -- may be the most difficult to refute in any definitive sense. The reasons are two. First of all, we rarely routinely collect direct measures of the "attractiveness" of the teacher to the students. Hence, indicators of this construct cannot be entered into our models and tested directly. The effect of "attractiveness" must, therefore, be inferred from the pattern of relationships among the other variables. However, the second problem in determining the effect of attractiveness on students' evaluations is that one would expect someone who provides "good service for good value" to become "attractive" to the consumer. Hence, it seems that one should not expect that "attractiveness" per se is a "contaminant" of students' ratings. Rather, the association between students' ratings and perceived attractiveness of the instructor should only be considered suspect when it can be demonstrated that the basis of the

attractiveness is something other than the qualities that make an individual a good teacher, or when it can be demonstrated that the perceived attractiveness of the instructor produces a "halo effect" that distorts the students' ratings.

Without providing a direct test of the "Attractiveness Hypothesis" here(9), examination of the differential associations among the variables in the present model can be used as an indicator that the variance among students' evaluations of different insttructors in different classes is not due primarily to halo effects. If it were, we would not expect to find that the "usefulness" of the course predicts quite differently to the two distinct ratings of course and instructor, or that the instructor's perceived patience and preparedness predict only to the overall ratings of the instructor and not to the students' ratings of the course.

Having disposed of much of the argument behind the "Structural Determinants" hypothesis by examining the (absence of) paths connecting most of the exogenous variables, even indirectly, with the students' overall ratings, let us turn to consider what evidence there is in support of the alternative "Effective Student Evaluator" hypothesis.

The first thing that might be noted in support of the "Effective Student Evaluator" hypothesis is the relatively large proportion of variance in each of the overall ratings accounted for by the component evaluations. We saw these strong associations in the bivariate associations between the components and the overall ratings (Table 4). Here, however, we are looking at a direct test of the hypothesis, stated as an implication of the alternative "structural determinants" view, that those bivariate relationships are spurious. The model in Figure 2 clearly indicates that this is not the case. Having removed all but the last remnants of the "structural determinants," we can still account for 68 percent of the variance in the students' overall ratings of the instructors and 57 percent of the variance in their overall ratings of the courses. And, it is clear from the coefficients given in Figure 2 that the component evaluations, based on the perceptions of the students, are all more important determinants of the overall evaluations than are the exogenous, structural variables. Further, the variables remaining in the model have reduced the correlation of .71 between the students' evaluations of the instructor and the course to a partial correlation, or correlation of the residuals, of .34. Squaring these values, we can see that the variables in the model have reduced the overlap between the students' two summary evaluations from 50 percent to about 12 percent, mostly due to the effects of the component evaluations.

Beyond the size of the effects from the component evaluations and the proportion of variance accounted for in the overall evaluations, the pattern of effects tend to support the "Effective Student Evaluator" hypothesis. The two factors most closely associated with differences in the personal style and approach of different instructors (i.e., their "preparedness" and "patience") evidence marginal effects only on the students' overall ratings of the instructor. On the other hand, the more general factors (i.e., "usefulness of the course" and "skillfulness of the instructor") impact on both the overall ratings of the instructor and the overall ratings of the course. If students are discriminating evaluators and unbiased judges of the quality of instruction, we ought to expect to find this kind of differentiation between the more "personal" characteristics of different instructors and the more "general" characteristics of classroom presentations.

On the basis of these findings, it seems apparent that students can and do discriminate distinct characteristics of their instructors and the experiences they have in different classes. It also appears clear that students use these discriminations in formulating their overall evaluations of their coursesand instructors, and that these overall evaluations are not carbon (or, more modern "xerox") copies of one another. In other words, the weight of the evidence based on the present analysis is in favor of the "Effective Student Evaluator" hypothesis and rather strongly against the "Structural Determinants" hypothesis.

But wait. What about the apparent effect of students' perceptions of the adequacies (or inadequacies) of the instructor's speech? Isn't this one of the "structural determinants" suggested earlier? Don't the data rather strongly support the idea that students are considerably biased against instructors they perceive as inadequate in communicating in the classroom?

Taking a second look at the model, one has to be struck by the strong relationships between the students' perceptions of the speaking characteristics of the instructor and each of the three most proximal components of the students' evaluations. The model indicates that instructors who are perceived as more adequate in their speaking ability are seen as better prepared for each class, more patient with students in the class, and genrally more skilled in their presentations of their courses. Each of these factors, in turn, affects the overall ratings students give to instructors and courses. And, still there is a direct effect of speech adequacies (or inadequacies) on the students' overall evaluations of the instructor. In fact, the total effect (i.e., direct plus indirect) of this variable on the overall rating of the instructor is larger than that of any other variable. Doesn't this say that the perceived speaking

characteristics of the instructor, considered as a "contaminant," is a strong predictor of the students' overall evaluations and support the structuralists' arguments?

It might sound petty at this point to disagree with such an impassioned arugument, but not known to be one to graciously avoid sounding petty, I will in fact disagree.

There are at least a couple of arguments against the structuralists' interpetation. First, it is possible that the model is misspecified, and students do, in fact, rate the instructor's speaking characteristics to reflect their predetermined evaluations of other aspects of the instructor's teaching. If this is the case, then such descriptions of the speaking characteristics of the instructor would not function as "contaminants" of the overall evaluations but simply as covariates.

If the original specification of the model is accurate, however, differential perceptions of speaking ability can, and apparently do, influence the students' evaluations. But, are these influences "contaminating"? To argue that they are, the structuralists must maintain either (a) that students are not competent judges even of whom they can better understand, or (b) that being able to communicate one's ideas to students in an understandable way is not relevant to good teaching. There is no evidence in the literature or in the present study to support the first contention, and the second seems to me absurd.

But, what if these data and the current interpretation on "speech effects" are accurate? If this is the case, there seems to be some hope. Speech patterns are manipulable, with training. If one works at it, one can improve one's methods of oral communication. (If this is not true, then why do we require undergraduates and high school students to take courses in speech communication?) The key, however, is "work." Teaching is a hard job. Teaching well over a long period of time to very heterogeneous and changing populations of students sometimes seems to be an impossible job. However, if the reports of the students in our classrooms are accurate, those who speak well tend to teach better. Maybe what the students are saying is that we need to pay more attention to this fundamental fact and do something to help those who do not communicate well with students improve their communication skills.

TEACHING AND ADMINISTRATIVE IMPLICATIONS OF THESE RESULTS

I have taken a rather light-hearted view of the various hypotheses faculty have to explain differences in students' evaluations. In ending this paper, I would like to somewhat more serious. Faculty are motivated to dismiss students' evaluations of our teaching for numerous reasons.

Some of these are legitimate. If it were, indeed, the case that students evaluated teaching performances and classroom experiences in a capricious manner or lacked the requisite knowledge and insights necessary to provide useful information, it would be our responsibility as educators and scientists to resist the use of these evaluations as an irrelevant political tool. If, however, it can be demonstrated that students have the requisite knowledge and skill to provide useful information, it behooves us both as educators and scientists to use this information to improve our teaching skills and performances.

I think the evidence in this paper points to students' evaluations as a useful tool in diagnosing the problems in our classrooms. Analyses of this type, if carried out conscientiously and with as much seriousness as we apply to our study of numerous other social processes, should alert individual faculty and administrators to the kinds of changes that will significantly improve our teaching and allow us to better educate our students.

I have just two prescriptions on the basis of the analysis reported above. My first prescription is that faculty ought to take seriously the theories they have regarding effective teaching and reject those self-serving hypotheses that justify our complacent attitudes toward our classroom activities. Further testing is necessary before we should definitively reject all the structural determinants of students' evaluations as ineffectual. However, I think such tests ought to be carried out and the results publicized so we, as a community of educators and not just as individuals, can benefit from some well-grounded theories and empirical tests. If we know what does not influence students' perceptions and evaluations of our teaching, we may be able to better focus on those factors that do influence the effectiveness of our teaching, and we can begin to do something about those factors.

My second prescription is to administrators and to faculty colleagues who act as administrators in our peer reviews of teaching. The analysis presented above, and the literature on which it is based, indicate that specific kinds of perceptions influence students' evaluations. A subsequent part of the analysis I have completed but have not reported here indicates that these perceptions are differentially distributed across faculty and account for almost all the variance in students' evaluations. By coding individual faculty in as "dummy variables" (no pun intended), we can determine who among our faculty are perceived in different ways by the students in their classrooms and how these perceptions influence the evaluations they receive from their students. In other words, we can use students' evaluations not only as the basis for our evaluation of teaching performances; we can also use the

students' perceptions to diagnose what individual instructors are doing that might be influencing these perceptions. On a somewhat more positive note, we can also identify those individuals who are perceived by their students to be doing an outstanding job, and -- if we are serious about improving the teaching quality of our programs -- we can match up the poorer teachers with the stronger ones, dimension by dimension, to bring assistance to those who most obviously need it.

The analyses above indicate that students' evaluations provide a credible basis for beginning the process of improving our teaching. They provide us the opportunity. If we take these evaluations seriously, we can be in a better position as individuals and colleagues to increase the educational benefits to our students and, I think, to improve the morale of our faculty.

TABLES

Table 1. Items Defining Six Factors Describing Instructors' Presentations of their Classes

N = 3589

	FACTORS	LOADINGS
FACTOR ONE: PREPARED		
Q6	Used class time well.	.816
Q9	Was well prepared for each class.	.793
FACTOR TWO: SKILLED		--------
Q7	Presented material in a logical and organized fashion.	.786
Q8	Gave adequate examples to clarify material.	.747
Q23	The instructor summarized and emphasized major points in lectures or discussions.	.724
Q19	The instructor clealy explained objectives of the course.	.682
Q22	Lectures and texts complemented each other.	.661
Q10	Used instructional materials where appropriate (blackboard, overhead projector, etc.).	.610
FACTOR THREE: ATTITUDE		--------
Q11	Presented material enthusiastically.	1.000 *

Continuation of Table 1.

FACTOR FOUR: PATIENCE		
Q12	Was patient with questions during class.	.859
Q21	Students felt free to ask questions and express their opinions.	.772
FACTOR FIVE: WORK		
Q15	Presented too much material for the available time.	.526
Q31	The amount of material covered was (a) excessive (b) above average (c) average (d) below average (e) too little.	.833
Q32	The amount of work required was (a) excessive (b) above average (c) average (d) below average (e) too little.	.569
FACTOR SIX: FAIRNESS		
Q25	The grading in the course was done fairly.	.730
Q20	Course requirements and grading system were clearly established.	.676
Q24	Tests included questions covering a broad sample of the presented material.	.636

Note: The single item indicator of Factor Three was set to be equivalent to the factor.

Table 2. Correlations Among Factors Describing Instructors' Presentations of their Classes

N = 3589

	PREPARED	SKILLED	ATTITUDE	PATIENCE	WORK	FAIRNESS
PREPARED	1.000					
SKILLED	.899	1.000				
ATTITUDE	.567	.582	1.000			
PATIENCE	.569	.682	.558	1.000		
WORK	.000	- .160	.015	- .150	1.000	
FAIRNESS	.656	.797	.447	.623	- .256	1.000

Table 3. Items Defining Three Factors Derived from Confirmatory Factor Analysis to Describe Students' Self-reported Responses to the Course

N = 3589

	FACTORS	LOADINGS
FACTOR ONE: USEFUL		
Q17	Feel better prepared to read and talk about materials in the area covered by the course.	.846
Q16	Found the course material useful for career preparation.	.765
Q14	Illustrated a way of thinking about material I found useful.	.701
FACTOR TWO: INTEREST INCREASED		
Q18	Found my interest in the subject matter was increased.	1.000 *
FACTOR THREE: DIFFICULTY OF COURSE		
Q33	For my background and ability, this course was (4) much too difficult (3) somewhat difficult (2) about right (1) somewhat easy (0) very easy.	1.000 *

Note: The single item indicators of Factors Two and Three were set to be equivalent to their respective factors.

Table 4. Correlations Among Factors Describing Students' Responses to their Classes

N = 3589

	USEFUL	INTEREST	DIFFCULT
USEFUL	1.000		
INTEREST	0.876	1.000	
DIFFCULT	-0.152	-0.130	1.000

Table 5. Bivariate Correlations of the Students' Component Evaluations with their Summary Ratings of the Instructor and the Course

N = 3589

	INSTRUCTOR	COURSE
INSTRUCTOR	1.000	
COURSE	.708	1.000
PREPARED	.774	.616
SKILLED	.794	.652
ATTITUDE	.570	.418
PATIENCE	.652	.510
WORK	-.070	-.126
FAIRNESS	.694	.578
USEFUL	.654	.720
INTEREST	.588	.662
DIFFICULT	-.062	-.115

Table 7. Correlations of the Structural Determinants with the Students' Summary Ratings of the Instructor and the Course

N = 3589

	INSTRUCTOR	COURSE
STATUS	.032	.021
MAJOR	.020	.018
LEVEL	.067	.070
REQUIRED	-.032	-.073
GPA	-.030	-.012
EXPGRADE	.166	.212
THEORY	-.028	-.114
METHODS	.021	.005
INSTORG	.036	.076
SOCPSYCH	-.024	-.008
CRIMJUST	.065	.094
NRATERS	.023	.020
PERMAJOR	.008	-.013
TERM	.000	.004
YEAR	.045	.023
SPEECH	.582	.417

Table 7. Descriptive Statistics for Scales and Items

N = 3589

VARIABLE	MEAN	STD DEV	DEFINITION
INSTRUCTOR	2.559	1.062	Student's rating of the instructor
COURSE	2.318	.896	Student's rating of the course
PREPARED	3.209	.790	SCALE1: Perceived preparedness
SKILLED	3.778	.747	SCALE2: Perceived teaching skill
ATTITUDE	3.052	1.069	SCALE3: Perceived enthusiasm
PATIENCE	3.543	.829	SCALE4: Perceived patience
WORK	1.264	.543	SCALE5: Perceived required workload
FAIRNESS	2.645	.729	SCALE6: Perceived fairness
USEFUL	2.097	.833	SCALE7: Usefulness of the course
INTEREST	2.615	1.192	SCALE8: Interest in the subject
DIFFCULT	2.213	.762	SCALE9: Difficulty with course
SPEECH	2.888	1.202	Adequacy of the instructor's speech
STATUS	2.714	1.017	Student's status (freshman, etc.)
MAJOR	.297	.457	Class in student's major? (Yes=0)
LEVEL	2.206	1.102	Level of course (freshman, etc.)
REQUIRED	.560	.496	Course required? (No=0; Yes=1)
GPA	2.658	.772	Reported grade point average
EXPGRADE	3.017	.696	Reported expected grade in course
THEORY	.068	.252	Theory course (dummy variable)
METHODS	.061	.239	Methods course (Dummy variable)
INSTORG	.208	.406	Institutions/organizations (dummy)
SOCPSYCH	.013	.114	Social psychology course (dummy)
CRIMJUST	.072	.259	Criminal justice course (dummy)
NRATERS	36.522	15.293	Number of students rating the course
PERMAJOR	30.388	25.389	% of class who were sociology majors
TERM	1.480	.500	Semester (Fall=1; Spring=2)
YEAR	82.436	.947	Fall, 1981 through Spring, 1984

Note: All scales and items, except TERM and YEAR and the dummy variables are coded on a scale of 0 to 4.00 to correspond to a 4-point academic grading scale.

PAGE RESERVED FOR FIGURE 1.

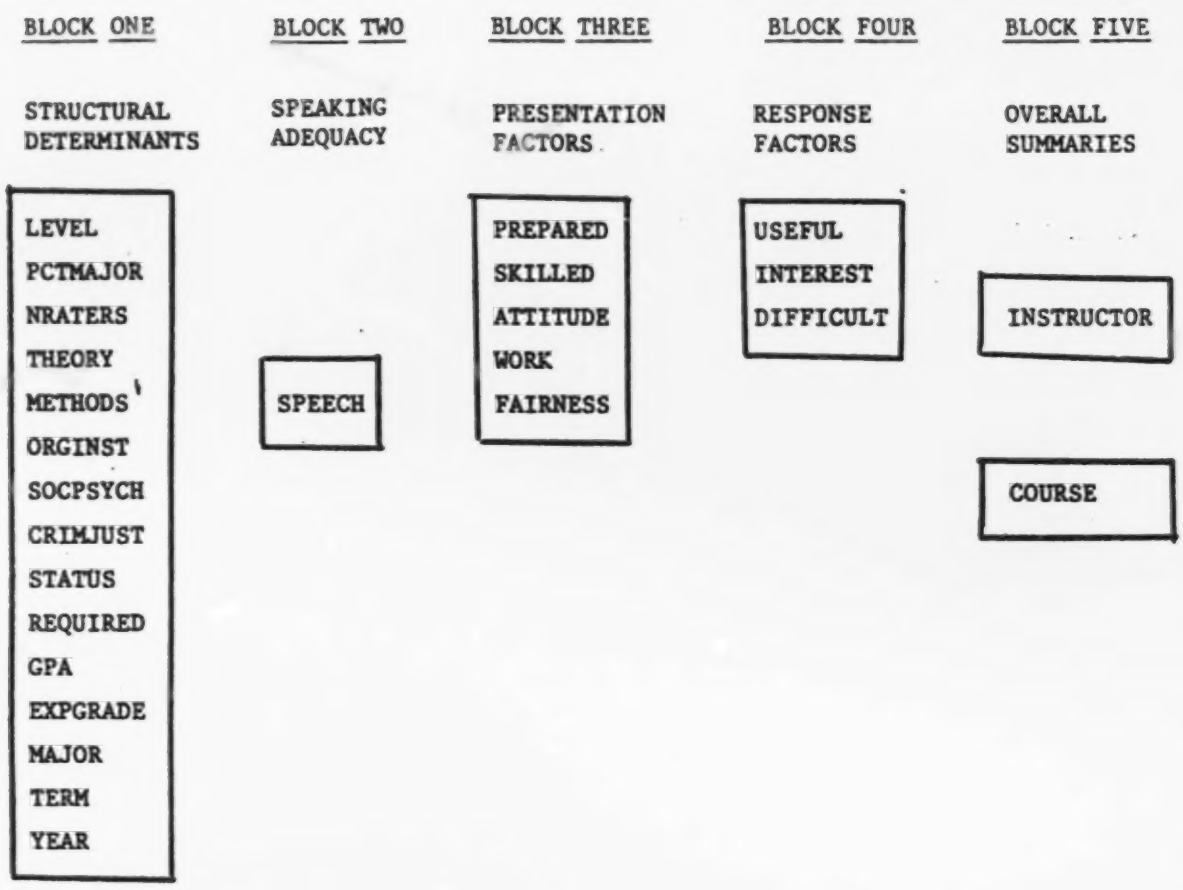

Figure 1. Specification for a Block Recursive Model

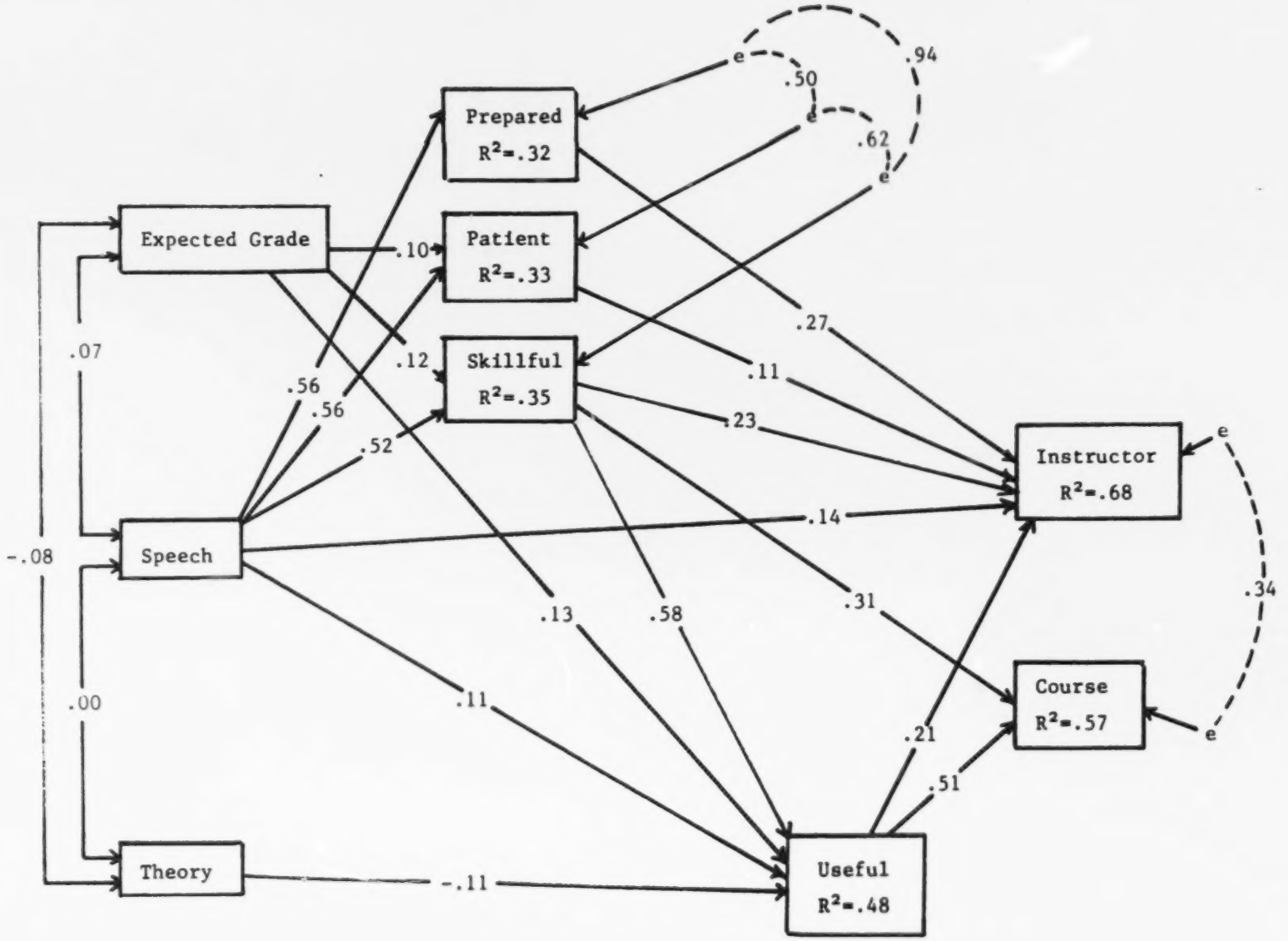

Figure 2. Block Recursive Path Model of Factors Influencing College Students' Ratings of Sociology Instructors and Courses (N = 3,589)

PAGE RESERVED FOR FIGURE 2.

NOTES

1. The exact wording of items used in the following analysis appear in the Appendix.

2. I wish to thank Dr. Herbert J. Rubin who was primarily responsible for the development of the instrument reported in this paper and the faculty of the Department of Sociology of Northern Illinois University who, over the course of many years, have provided abundant input and comment about the instrument and the students' responses.

3. The Goodness-of-Fit Index measures the total proportion of variance and covariance accounted for in the model. The Adjusted Goodness-of-Fit Index measures the total proportion of variance and covariance accounted for in the model adjusting for the number of degrees of freedom used to fit the data. The Chi-square Goodness-of-Fit of 1898.14 with 105 degrees of freedom was not used as a criterion for acceptance of the model given the very large sample size and the fact that no attempt was being made in this preliminary analysis to take into account the complexity of the variables or to fit the correlations among the error terms. For simplicity of conceptualization adn presentation, a standard factor model was specified with uncorrelated errors, and complete simple structure was maintained to facilitate interpretation of the constructs as each associated entirely with the covariance among the items loading on each factor. Subsequent analyses are anticipated in which it is expected, given the present results, that a somewhat different factoral structure will be specified combining two or more of the present, highly intercorrelated, factors, and at least one additional factor will be defined consistent with high correlations between pairs of items loading on the same and different factors. Crossvalidation of this new specification will be done using the respondents selected out of the present analysis.

4. The Chi-square Goodness-of-Fit value is 58.09 with 4 degrees of freedom.

5. A number of variables also implied by the "alternative hypotheses" have not been coded for all N= 3589 students in the present analysis. These include: actual enrollment in the class, time of day, day of week, length of class (in minutes), and the response rate on the evaluations. Previous analysis of these data (Suchner, 1984) and previous research by others (cf., Feldman,

1979) have demonstrated that none of these variables have any appreciable effect on students' overall evaluations, or much effect on any of the components, once the other factors already included in the model are controlled.

6. It is the second of these assumptions that appears to be most debatable. To the extent that students infer the adequacy of an instructor's speech from other aspects of his/her presentation of the the course, this specification will attribute causal impact to the instructor's speech characteristics rather than to the students' other perceptions.

7. Eleven percent of the respondents were in classes with 20 or fewer raters in the present data, while nine percent were in classes with more than 60 raters. Given that an analyses (Suchner, 1984) on a smaller subset of the present data (N = 2743) indicated an average response rate of 62.5 percent, the estimated class sizes for the eleven percent in the smallest classes is a maximum of 32 students, while the estimated class size for the nine percent in the largest classes is a minimum of 96 students. Hence, there appears to be enough variance in class sizes to have detected the hypothesized curvilinear relationship if it, indeed, existed.

8. With a sample size of N = 3589, numerous very small path coefficients were statistically reliable. Hence, the criterion of "substantive significance" was employed to remove paths with coefficients less than .10.

9. The author, with the help of his colleagues in the department, has constructed a direct test of the "Attraction Hypothesis" through inclusion of items in the Fall, 1984, student evaluation questionnaire designed to measure the differential "attractiveness" of seven different instructors in 14 different classes. I will report the findings of this test in the final version of this paper, if I have the opportunity to complete the analysis prior to the meetings.

REFERENCES

Braskamp, L. A., Caulley, D., & Costin, F. Student ratings and insrtuctor self-ratings and their relationship to student achievement. American Educational Research Journal, 1979, 16, 295-303.

Centra, J. A. Student ratings of instruction and their relationship to student learning. Research Bulletin 76-6. Princeton: Educational Testing Service, 1976.

Costin, F. "Intercorrelation between Students' and Course Chairmen's Ratings of Instructors." University of Illinois, Division of General Studies, 1966. (mimeographed.)

Costin, F. A graduate course in the teaching of psychology: description and evaluation. Journal of Teacher Education. 19, 1968, 425-432.

Costin, W. T., Greenough, W. T., and Menges, R. J. Student ratings of college teaching: reliability, validity and usefulness. Review of Educational Research. 41, 1971, 511-535.

Feldman, K. A. Consistency and variability among college students in rating their teachers and courses: A review and analysis. Research in Higher Education, 1977, 6, 223-274.

Feldman, K. A. The significance of circumstances for college students' ratings of their teachers and courses. Research in Higher Education, 1979, 10, 149-172.

Feldman, K. A. The superior college teacher from the students' view. Research in Higher Education, 1976, 5, 243-288.

Gaff, J. Teachers who make a difference. Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students, ed. A. Sockcoff. Philadelphia: Temple University, 1973.

Gage, N. L. The appraisal of college teaching. Journal of Higher Education, 32, 1961, 17-22.

Goodhartz, A. S. Student attitudes and opinions relating to teaching at Brooklyn College. School and Society. 68, 1948, 345-349.

Hamilton, L. C. Grades, class size, and faculty status predict teaching evaluations. Teaching Sociology, 1980, 8, 47-60.

Harvey, J. N. and Barker, D. G. Student evaluation of teaching effectiveness. Improving College and University Teaching. 18, 1970, 275-278.

Heider, F. The Psychology of Interpersonal Relations. New York: Wiley, 1958.

Hildebrand, M., Wilson, R. C., and Dienst, E. R. Evaluating University Teaching. Berkeley: University of California, Center for Research and Development in Higher Education, 1971.

Hoffman, R. G. Variables affecting university student ratings of instructor behavior. American Educational Research Journal, 1978, 15, 287-299.

Joreskog, K. G. and Sorbom, D. LISREL VI: Analysis of Linear Structural Relationships by the method of maximum likelihood. Mooresville, IN: Scientific Software, 1984.

Kohlan, R. G. A comparison of faculty evaluations early and late in the course. Journal of Higher Education. 44, 1973, 587-594.

Lovell, G. D., & C. F. Haner. Forced-choice applied to college faculty rating. Educational and Psychology Measurement, 15, 1955, 291-304.

Marsh, H. W. Factors affecting students' evaluations of the same course taught by the same instructor on different occasions. American Educational Research Journal, 1982, 19, 485-497.

Marsh, H. W. The influence of student, course, and instructor characteristics in evaluations of university teaching. American Educational Research Journal, 1980, 17, 219-238.

Marsh, H. W. and Overall, J. U. The relative influence,1981, of course level, course type, and instructor on students' evaluations of college teaching. American Educational Research Journal, 18, 103-112.

Marsh, H. W., & Hocevar, D The factorial invariance of student evaluations of college teaching. American Educational Research Journal, 1984, 21, 341-366.

Marsh, H. W., Overall, J. U., & Kesler, S. P. Class size, students' evaluations, and instructional effectiveenss. American Educational Research Journal, 1979, 16, 57-69.

Maslow, A. H. and Zimmerman, W. College teaching ability, scholarship, activity and personality. Journal of Educational Psychology, 47, 1956, 185-189.

McDaniel, E. D. and Feldhusen, J. F. Relationships between faculty ratings and indexes of service and scholarship. Proceedings of the 78th annual convention of the American Psychological Association. 5, 1970, 619-620.

Miller, R. I. Assessing teacher effectiveness. Proceedings of the First International Conference on Improving University Teaching, ed. B. Massey. Heidelberg: University of Maryland, 1975.

Miller, R. I. Developing Programs for Faculty Evaluation. San Francisco: Jossey-Bass, 1974.

Miller, R. I. Evaluating Faculty Performance. San Francisco: Jossey-Bass, 1972.

Murray, H. G. The reliability and validity of student ratings of faculty teaching ability. Unpublished manuscript, 1973.

Remmers, H. H. On Students' Perceptions of Teachers' Effectiveness. The Appraisal of Teaching in Large Universities, ed. W. J. McKeachie. Ann Arbor: University of Michigan Extension Service, 1959.

Rodin, M. Student evaluation. Science, 1975, 187, 555-557.

Seldin, P. Successful Faculty Evaluatuion Programs. Crugers, NY: Coventry Press, 1980.

Solomon, D. Teacher behavior dimensions, course characteristics, and student evaluation of teachers. American Educational Research Journal, 3, 1966, 35-47.

Spencer, R. E. The Illinois course evaluation questionnaire: manual of interpretation. Reserch report No. 270. Champaign: University of Illinois, 1968.

Spencer, R. E. and Aleamoni, L. M. A student course evaluation questionnaire. Journal of Educational Measurement, 7, 1970, 209-210.

SPSS, Inc. SPSSX User's Guide. New York: McGraw-Hill, 1983.

Suchner, R. W. Teaching Sociology: What Dimensions Do Students Discriminate? Paper presented to the Illinois

Sociological Association annual meetings, Chicago, Illinois, October 19, 1984.

Toug, M. S. and Feldhusen, T. F. The relationship between student ratings of instructors and their participation in classroom discussion. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, February, 1973.

Toug, M. S., Feldhusen, T. F., and Halstead, J. Criterion-reference validity of student ratings. Paper read at annual meeting of the American Educational Research Association, 1973.

Wimberley, R. C., Faulkner, G. L., & Moxley, R. L. Dimensions of teaching effectiveness. Teaching Sociology, 1978, 6, 7-20.

APPENDIX

INSTRUCTIONS AND WORDING OF THE ITEMS USED IN THE PRESENT ANALYSIS

Coded values (in parentheses) did not appear on the questionnaire. The items are labelled to correspond to their original order in the questionnaire. The variable names that appear in this paper are inserted in parentheses following the item number.

EVALUATION INSTRUMENT FOR SOCIOLOGY

University policy requires that a student evaluation of instruction be conducted in this class. The results of this evaluation will be considered by the departmental personnel committee in the process of determining merit ratings, which in turn will affect decisions concerning salary, promotion, and tenure. Your instructor will also consider the results of this questionnaire in future course preparation.

Please complete this evaluation in a responsible and thoughtful manner. To ensure your uninhibited expression, you are not required to sign your evaluation. The results will be distributed to faculty members only after the course is completed and grades have been submitted to the registrar.

I. General Information

1. (STATUS) Class standing: (1) freshman (2) sophomore (3) junior (4) senior (5) graduate

2. (MAJOR) Is this class in your major department? (0) no (1) yes

3. (REQUIRED) Course was (0) elected (1) required

4. (GPA) Your current G.P.A. is (1) 1.99 or below (2) 2.00-2.75 (3) 2.76-3.25 (4) 3.26-4.00 (2.5) no G.P.A. yet--just started school

5. (EXPGRADE) Grade expected (4) A (3) B (2) C (1) D or F (2.5) taking course pass-fail

BELOW: (4) SA = strongly agree
(3) MA = mildly agree
(2) N = neither agree nor disagree

(1) MD = mildly disagree
(0) SD = strongly disagree

II. Teaching Skill

The instructor:

Q6 Used class time well.

Q7 Presented material in a logical and organized fashion.

Q8 Gave adequate examples to clarify material.

Q9 Was well prepared for each class.

Q10 Used instructional materials where appropriate (blackboard, overhead projector, etc.).

Q11 Presented material enthusiastically.

Q12 Was patient with questions during class.

Q13 Speech is adequate for teaching (volume, intelligibility, etc.).

Q14 Illustrated a way of thinking about material I found useful.

Q15 Presented too much material for the available time.

III. My Reaction

I myself

Q16 Found the course material useful for career preparation.

Q17 Feel better prepared to read and talk about materials in the area covered by the course.

Q18 Found my interest in the subject matter was increased.

IV. Course Structure

Q19 The instructor clealy explained objectives of the course.

Q20 Course requirements and grading system were clearly established.

Q21 Students felt free to ask questions and express their opinions.

Q22 Lectures and texts complemented each other.

Q23 The instructor summarized and emphasized major points in lectures or discussions.

V. Evaluation Quality

Q24 Tests included questions covering a broad sample of the presented material.

Q25 The grading in the course was done fairly.

VI. Other Comments and Summary Comments

Q31 The amount of material covered was (4) excessive (3) above average (2) average (1) below average (0) too little.

Q32 The amount of work required was (4) excessive (3) above average (2) average (1) below average (0) too little.

Q33 For my background and ability, this course was (4) much too difficult (3) somewhat difficult (2) about right (1) somewhat easy (0) very easy.

Q35 (INSTRUCTOR) Overall, I would rate the instructor (4) superior (3) above average (2) average (1) below average (0) poor.

Q36 (COURSE) Overall, I would rate the course (4) superior (3) above average (2) average (1) below average (0) poor.

ANY OTHER COMMENTS YOU MIGHT LIKE TO MAKE WILL BE APPRECIATED. PLEASE WRITE SUCH COMMENTS ON THE BACK OF THIS SHEET AND RETURN THIS SHEET ALONG WITH THE IBM SHEET.

Some items asked on the original questionnaire were not included in the present analysis.